Improving protein structure prediction using petascale sequence search

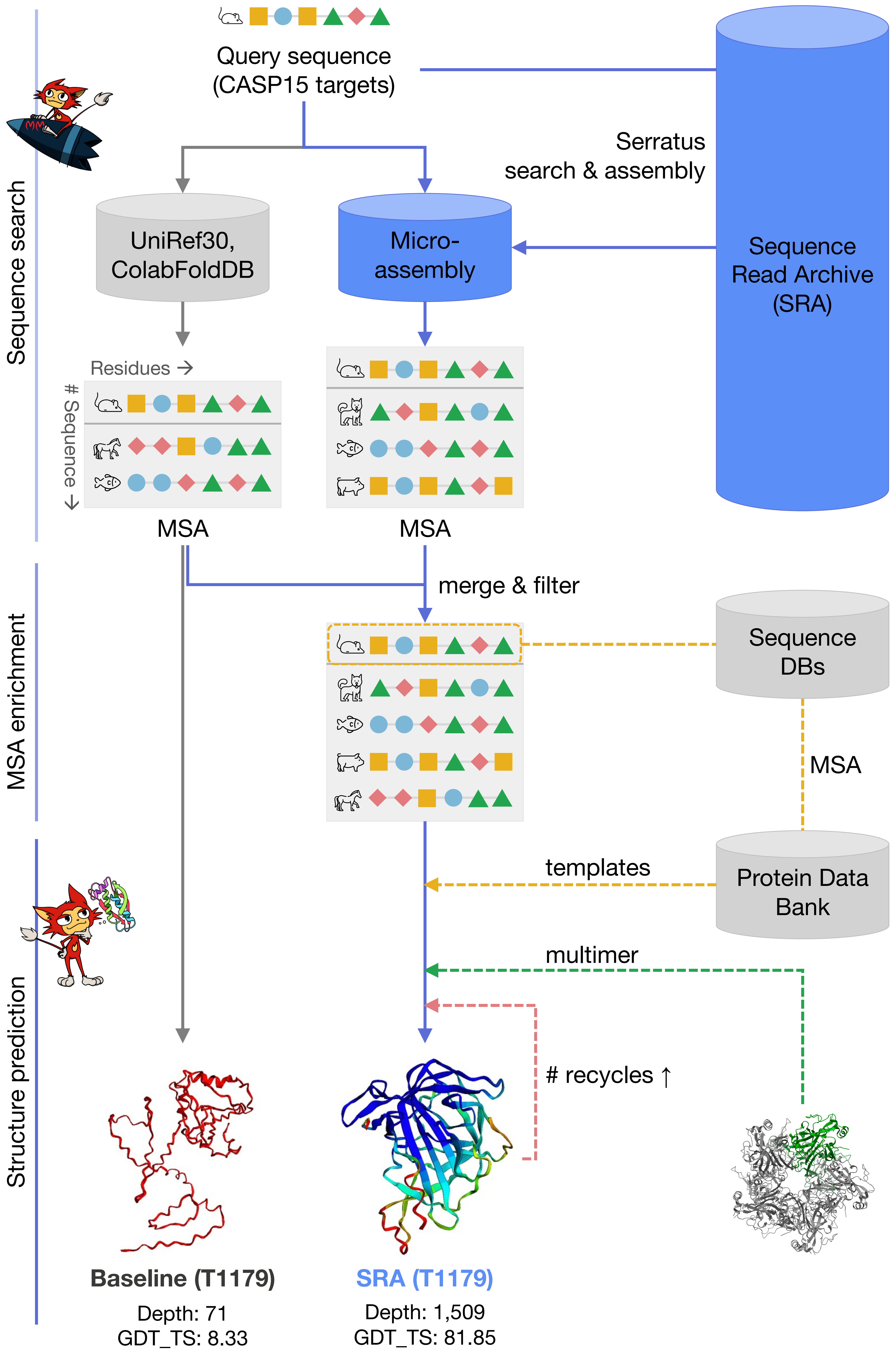
Sewon Lee[1,*], Gyuri Kim[2,*], Milot Mirdita[1], Eli Levy Karin[3], Sukhwan Park[1], Rayan Chikhi[4], Artem Babaian[5], Andriy Kryshtafovych[6] and Martin Steinegger[1,†]

[1]School of Biological Sciences, [2]School of Biology Education, Seoul National University, South Korea; [3]ELKMO, Denmark; [4]Department of Computational Biology, Institut Pasteur, France; [5]Independent researcher, Canada; [6]Genome Center, University of California, Davis, USA
*Equal contributors; †Correspondence: martin.steinegger@snu.ac.kr

Abstract

In the recent CASP15 competition, the crucial role of multiple sequence alignments (MSAs) in protein structure prediction was underscored by the success of AlphaFold2-based models, such as ColabFold. To push the boundaries of MSA utilization and understand its limits, we conducted an exhaustive, petabase-scale search of the Sequence Read Archive (SRA), the world's largest public sequence database, using CASP15 targets as queries. Utilizing ColabFold, an accelerated version of AlphaFold2 offering numerous advanced features, we merged the SRA- and the baseline MSAs to predict the structure of each query. This approach significantly improved structure prediction accuracy, achieving a high GDT_TS (>70) for 66% of the non-easy targets, a substantial leap from the 52% achieved with baseline ColabFold MSAs. Enabling ColabFold's advanced features, including more recycles, using templates, and multimer models, contributed to a further performance boost. This significant increase in accuracy improved ColabFold's CASP15 ranking from 11th to 3rd place among 47 server groups, indicating the vast potential of large-scale sequence exploration for better structure prediction.

Deep & wide sequence search for structure prediction



Baseline (T1179)
Depth: 71
GDT_TS: 8.33

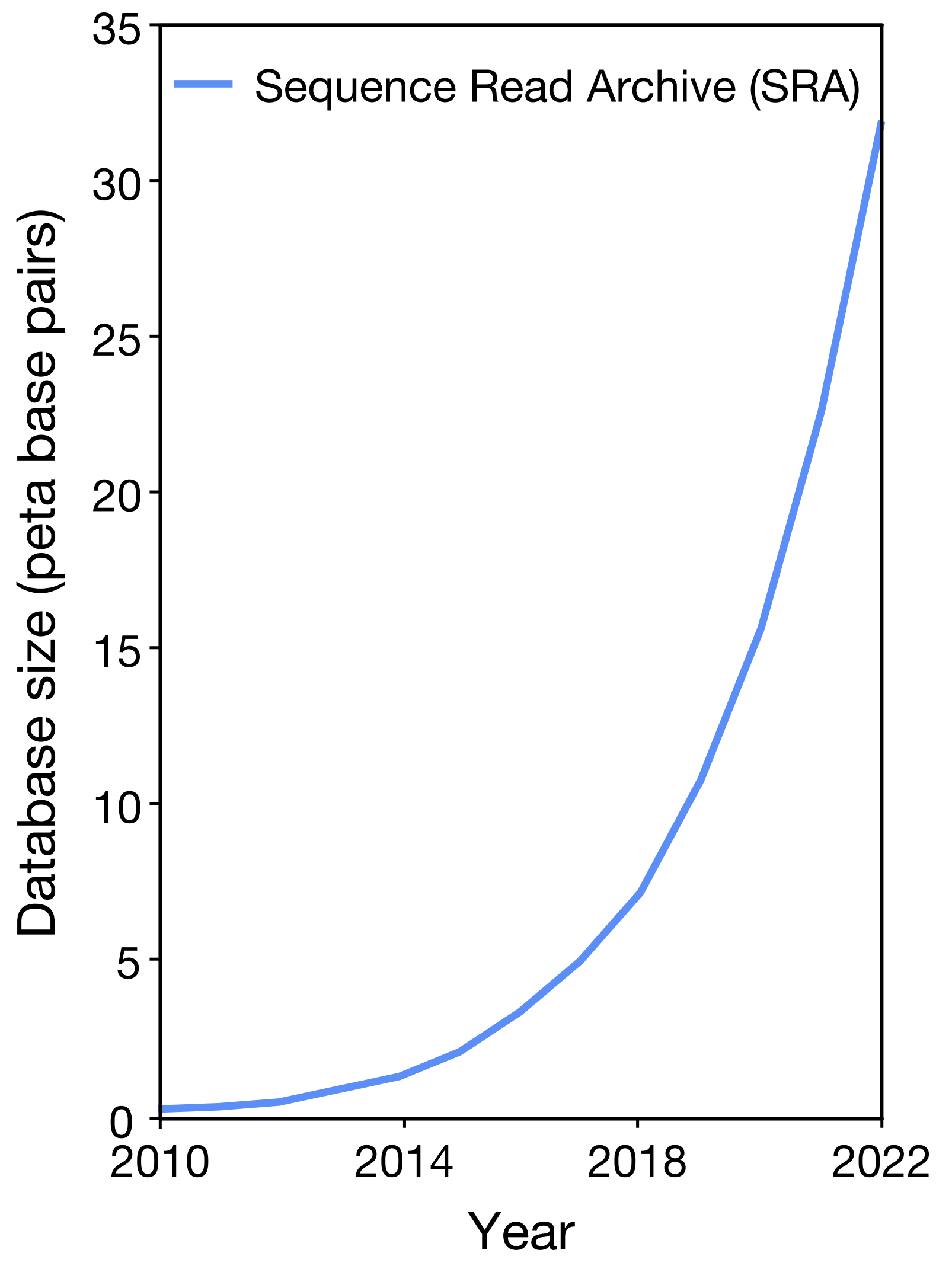
SRA (T1179)
Depth: 1,509
GDT_TS: 81.85

SRA: the largest public sequence database

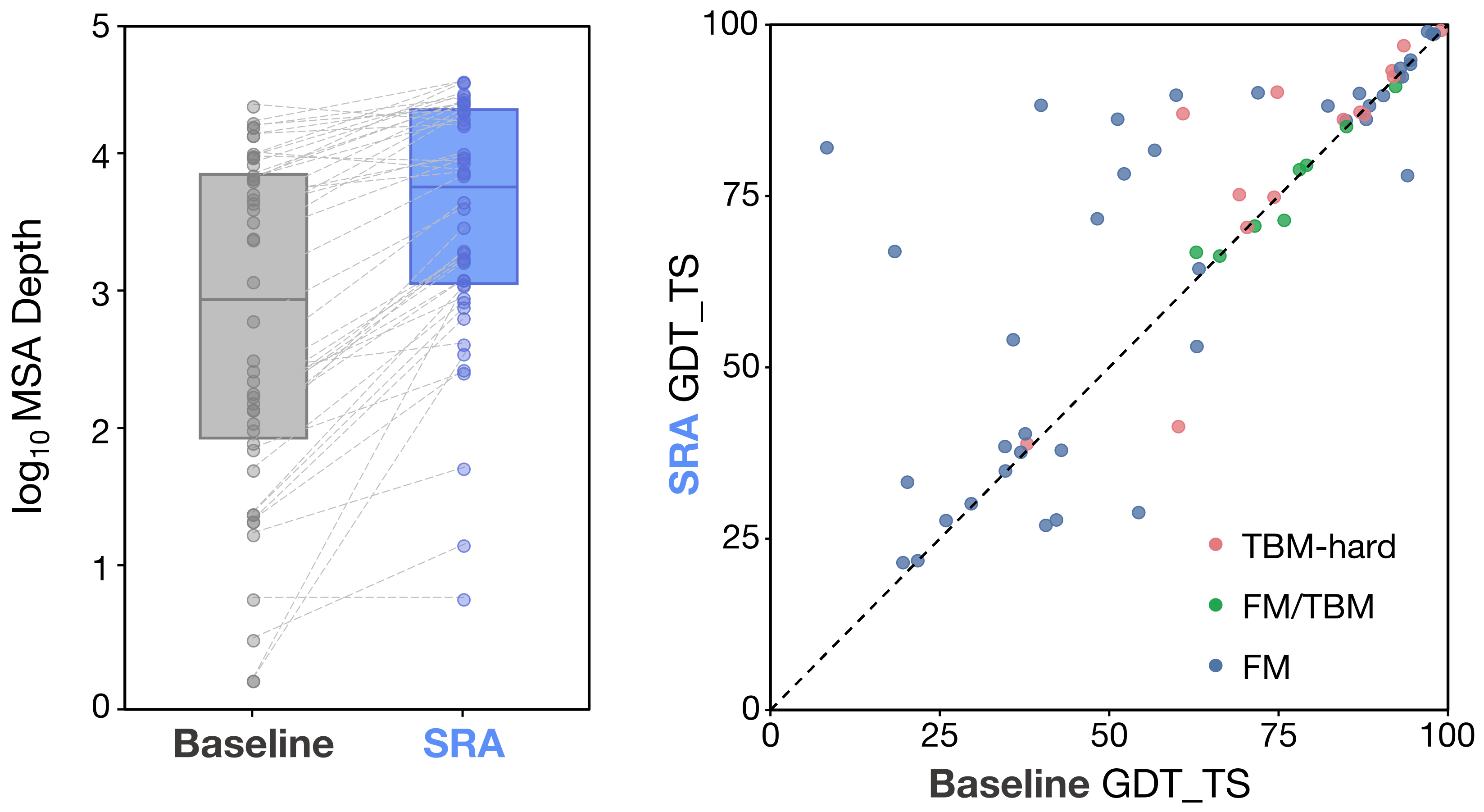
Database	Giga base pairs
SRA	32,000,000
IMG/M	27,000
BFD*	2,600
Metaclust*	1,700
Microassembly	**1,500**
ColabFoldDB*	800
MGnify*	300
UniRef30*	30

*Estimated based on number of sequences

Microassembly:
assembled *search results* from SRA

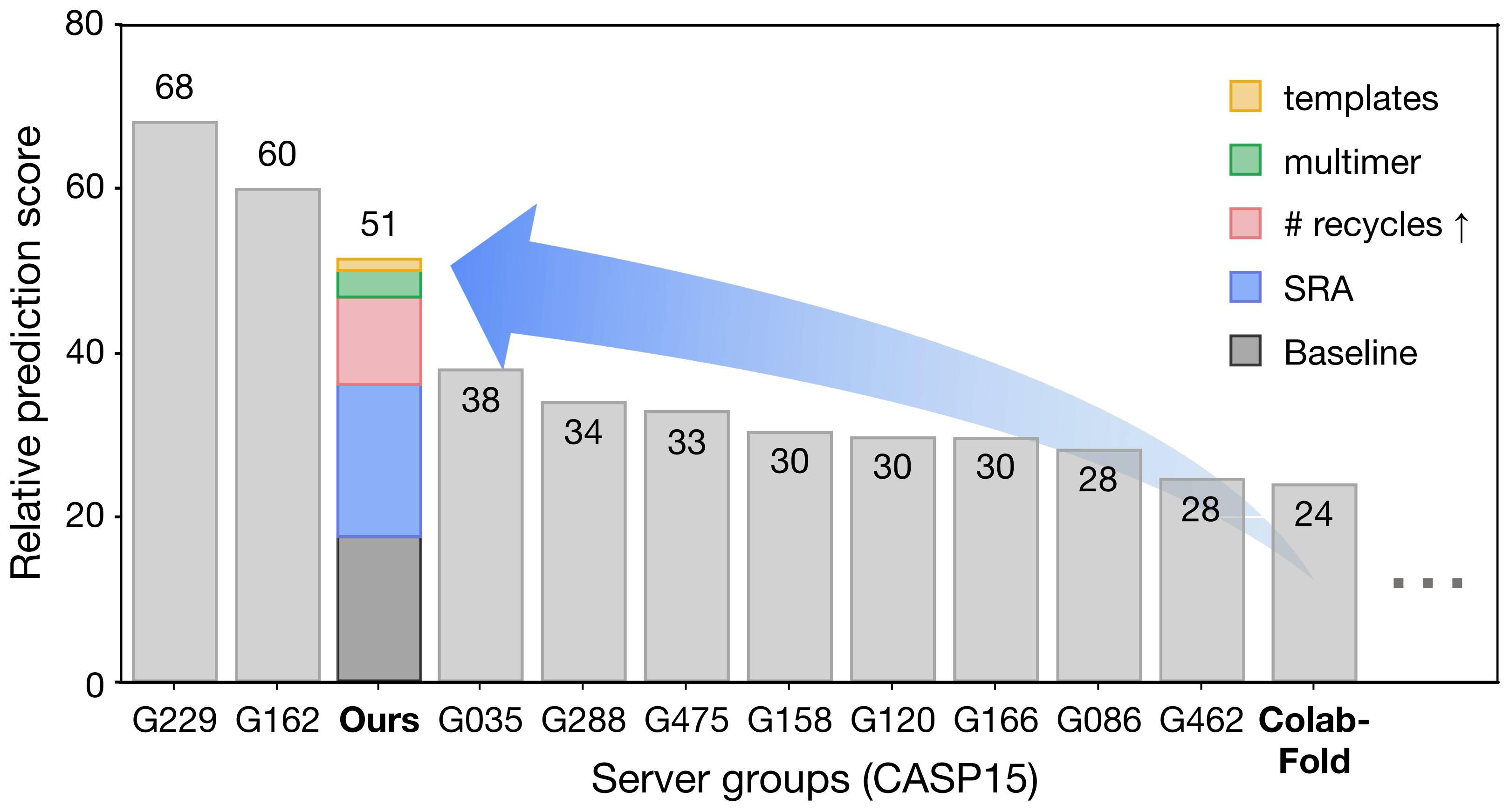


SRA-enriched MSAs lead to better predictions



TBM: templated-based modeling; **FM:** free-modeling
GDT_TS (global distance test): prediction accuracy

Factors contributing to improved prediction



Score: sum of GDT_TS (accuracy) Z-scores over 0
Contribution: SRA > # recycles ↑ (3 → 12) > multimer > templates
CASP15 rank: 11th → 3rd among 47 server groups

References

Kryshtafovych, A., et al. *Proteins*, *89,* 1607-1617 (2021).
Jumper, J., et al. *Nature*, *596*, 583-589 (2021).
Ovchinnikov, S., et al. *Science*, *355*, 294-298 (2017).
Mirdita, M., et al. *Nat. Methods*, *19*, 679-682 (2022).
Edgar, R. C., et al. *Nature*, *602*, 142-147 (2022).
Steinegger, M., & Söding, J. *Nat. Biotechnol.*, *35*, 1026-1028 (2017).

Download Poster (PDF)